Vocabulary Conversion: Performance with Controlled and Uncontrolled Terms and Tags

Technical Report Number TR-2008-02
School of Information
and Library Science
University of North Carolina
at Chapel Hill

Robert M. Losee*

January 6, 2009

Abstract

Controlled and uncontrolled indexing terminology and metadata may be converted from one to another. Decision criteria are developed that can be used to determine which terms should be assigned when converting vocabularies. Methods are developed for computing the parameters of these systems, as well as means for estimating the parameters when given limited information. These conversion techniques may be applied to thesaurus terminology, gene ontologies, topic maps, uncontrolled natural language terms, folksonomies, tags and labels on web pages, the presence or absence of a specific hyperlink, as well as to metadata. Rules are provided suggesting circumstances when controlled vocabularies are always superior to using uncontrolled vocabularies.

1 Introduction

Different people may use different terminology when referring to the same thing. While adult native speakers of English might use the term *dog* to refer to

*Thanks to Earl Bailey and Lee Roush for comments on this manuscript.

a barking canine, English speaking children often use the term *doggy*. Speakers of different languages represent the same animal sounds using different words, such as when an English speaker refers to a dog barking using *arf arf* while a Japanese speaker might use *wan wan*, a Chinese speaker *wang wang* and a Turkish speaker *hev hev*. Below, we address the following question: when does one decide to convert or translate from one user or system's vocabulary or metadata to the vocabulary used by another person or system? How can a vocabulary system be shown to be better than another for a particular user, a group, or an organization, making the terminology truly user-centered? Conversion from one vocabulary to another may occur when moving from one language to another, such as French to English, when merging databases with vocabulary from two different indexing systems, each with its own vocabulary, or when converting one direction or another between user supplied uncontrolled terms and controlled terminology. Some example calculations are provided in the second half of this article.

Vocabularies take many forms (Cleveland & Cleveland, 2001; Foskett, 1996; Furnas, 1985). At the core of all vocabularies are sets of *terms*, individual words or phrases (Aronoff & Fudeman, 2005). In computer applications, it is common to define a term as a single group of characters with no intervening space characters, whereas a phrase might be more than one such group of characters. For notational simplicity, we will refer below to a single word or a short phrase, such as a *house*, *house boat*, *Greco-Roman knuckle lock*, or *Claude E. Shannon*, as a vocabulary or metadata *term*.

A vocabulary is a set of terms that may be used to represent objects, ideas, and other phenomena in a particular domain. When developing a *controlled vocabulary*, the designers intend that a single controlled vocabulary term solely represents an idea, class, or object. By having a single vocabulary reference for each object, idea, or other phenomenon, all items on a specific topic can be accessed as a group. Because all documents on a topic will have the same vocabulary term as a label, use of the label will retrieve all documents on a given topic, providing for high recall retrieval, especially useful in small datasets or in large datasets in areas such as medicine or law, where the searcher wishes to find all the material on a given topic in order to satisfy their information need. Experiments have repeatedly shown that a well developed and accurately applied controlled vocabulary will produce better average retrieval performance than when using an uncontrolled vocabulary (Cleverdon, 1967; Foskett, 1996; Salton & Lesk, 1968). Note that this does not mean that controlled vocabularies always outperform uncontrolled vocabularies; a controlled vocabulary may be poorly developed or poorly applied by indexers or poorly used by searchers, producing weaker performance for the controlled vocabulary than for the uncontrolled vocabulary.

One finds that formally developed controlled vocabulary systems, such as those used in libraries, retrieval systems, and metadata applications, often use terminology that differs from the freely chosen uncontrolled terms selected by many of their patrons when the system provides users with unlimited vocabulary choices. Controlled vocabularies are found most frequently where organizations index large quantities of documents for end-users. A controlled vocabulary is often provided in an organization-produced thesaurus as a listing of hierarchically related terms, along with cross references between terms. Professionals using a thesaurus will index individual documents or media, bearing the expense of bringing together similar documents by once assigning index terms to a document, benefiting all the users of the organization's indexing services.

An uncontrolled vocabulary may be used when a system designer allows a searcher, manual indexer, or computer indexing software to use their own terminology to represent topics. System users, for example, may search using whatever terms they choose to use, without needing to refer to the list of "acceptable" controlled vocabulary terms. System designers do not need to develop or use a controlled vocabulary representation for each item, decreasing significantly the cost associated with providing access points for an item added to the collection. Many search engines, for example, choose to represent documents only with terms contained in the document text rather than with controlled vocabulary terms or phrases. These search programs usually support searching with any user-supplied terms, rather than with a smaller set of controlled terms. An author's choice of terms, as well as a searcher's choice of terms, may allow information to be produced and used by those not sharing the world views of those who designed a particular controlled vocabulary.

It may be desirable to convert vocabulary from one system to another. For example, one might wish to shift from one set of controlled terms to a different set of controlled terms, such as when one moves a set of documents from one organization's vocabulary to the vocabulary of another organization. One might move a set of documents indexed by the American Psychological Association, using their controlled vocabulary, to a system using the Library of Congress Subject Headings. One might wish to translate material from the Library of Congress Subject Headings to a set of uncontrolled terms.

While machine translation from one natural language to another is very difficult to perform, the translation or conversion from one subject system to another works with "cruder" (Maniez, 1997, p. 218) objects and is a far simpler task. Below we present performance criteria for translating vocabulary, such as when moving from an uncontrolled system to a controlled system improves performance, or when moving from one uncontrolled system such as folksonomies to another uncontrolled system improves performance. Expected

or past performance can serve as the basis for making decisions about vocabulary. Methods are presented that will allow one to make decisions, and the decision criteria may be understood both in terms of its capabilities and in terms of the assumptions consistent with the decision model.

2 Vocabularies

Information systems often contain terms as components to be used when computing the degree of similarity between two media fragments, such as a query and a document. In a structured (vs. non-structured) database, such as a simple telephone directory, one might look for an exact match when searching for a particular person's name in the database. In an unstructured database, however, one might look for the occurrence of the search term anywhere in the document, or one might search for a related term or morphological variant, such as the singular form when the search term is a plural form of the term. Terms can serve as perfect or imperfect indicators of the nature of the object they represent, and variants of terms and their contexts can carry differences in meanings (Bowker & Hawkins, 2006). Without perfect indicators, combining the similarities between imperfectly indicating terms will often improve the accuracy of classification or ranking procedures.

One can evaluate a vocabulary by measuring the performance of a system that uses this vocabulary (Cleverdon, 1967; Losee, 2001a, 2006b; Vivaldi & Rodriguez, 2007). The information retrieval system that is used as a testbed will usually have a database consisting of documents, queries, and relevance judgments, with document representations containing the vocabulary being studied. The performance using a system s is denoted as $\mathcal{P}(s)$. Asking whether the performance of a controlled vocabulary c exceeds the performance with a uncontrolled vocabulary u is denoted as asking if the inequality $\mathcal{P}(c) > \mathcal{P}(u)$ holds. One specific method of computing system performance values will be presented in a later section.

In decision making, one normally chooses to take the action associated with the largest expected value for payoff or performance. However, in some circumstances one must choose a system based on managerial or political concerns, regardless of its relative expected performance.

The use of vocabulary terms places text in a *context*, the meaning of terms often being determined in part by the text occurring around the terms. Describing a particular context may be understood as capturing the nature of an abstract idea. A context may be implemented as a set of terms that are viewed as characteristic of the context, with other entities that contain these terms being understood to be capable of being placed in the context.

Sets of contextual terms may be derived using several different methods (Church, Gale, Hanks, Hindle, & Moon, 1994; Grefenstette, 1992; Pearce, 2002; Weeds & Weir, 2005). The relationships between the different methods may best be seen by considering the work of Pearce (2002), which explores the performative relationships between collocation methods. One method is to find terms that tend to occur together, referred to as terms that *co-occur*. Beginning with a seed term, one may find other terms that have a strong degree of association with the seed term. These may be terms that mean the same thing and the synonyms are lexical substitutes, or they may be terms that address the same context but do not have the same meaning, such as the set of terms: *boat*, *river*, *paddle*, and *row*.

Collocated terms may also provide contextual information because topical discussions inherently include related terms. These include terms occurring in phrases such as *White House*, *family business*, *research university*, and *European Union*. Collocated terms do not necessarily have the same meaning, but because the terms occur together, they frequently address the same domain. For example, the term *school* might co-occur with *teacher*, *student*, and *classroom* because conversations about one often entail discussions of the others. Terms co-occurring in an uncontrolled vocabulary often represent a single term in a controlled vocabulary, e.g., *education*. The set of uncontrolled vocabulary terms that should be used together to produce a controlled term or terms may be determined by searching for those uncontrolled terms that have a high expected mutual information measure (EMIM) with a seed controlled term or a seed uncontrolled term (Church & Hanks, 1990; Li, 1989; Losee, 2001b).

3 Vocabulary Conversion, Translation, & Terminology in Different Languages

Terms used to index material in one indexing system may be converted to a different indexing system. The translation may be direct, from a source vocabulary to a destination vocabulary, or the translation may move from one vocabulary through an intermediate language or "switching language" (Maniez, 1997, p. 220) and then to the destination language. Using an intermediate language allows a translation system to be developed with the capability for converting from each possible source language to a single intermediate language, as well as a set of values for the conversion from the intermediate language to the destination languages. Using a single intermediate language is easier than developing capabilities for each of the possible conversions that might take place from each possible source to each possible destination language. This number of possible conversions is proportional to the product of the number

of source languages times the number of destination languages (Maniez, 1997). When converting by moving through an intermediate language, the number of conversions possible is much smaller, being computed as the number of possible conversions from the source to the intermediate language plus the number of conversions possible from the intermediate language to the destination language.

Vocabularies and terms are representations for mental constructs held or generated by one or more people. One person's cognitive structures may be similar in some respects to another person's structures, but likely differ in other respects. The vocabulary these two individuals use to represent identical or similar ideas may be the same for each person or may differ. The two people may use the term *purple* to describe the same underlying color, while one may use the term *red* to represent what the other would call *green*. There is clearly a social aspect to the effective use of vocabulary, and a person who constantly uses the term *red* to refer to what others refer to as *green* will have difficulty communicating with others, and will often be "corrected" or encouraged to use the terms found in the socially dominant idea–term mapping.

The same idea and minor variants of it may be expressed in different languages using different terms. Natural language translation of full text in one language to another can involve a variety of possible processes, consistent with different models of language and its underlying meaning. Many different grammars have been proposed by linguists, and the translation of natural language text will vary with the assumptions made by the different grammars and models of language (Chomsky, 1965; Dodd, Campbell, & Worrall, 1996; Manning & Schutze, 1999; Newmeyer, 1986). Most essential to the translation process is capturing the meaning expressed in one language, the source language and producing it in the destination language. This is often achieved by the individual translation of individual terms from one language to another. However, words often have multiple meanings, and need to be disambiguated through the examination of context. For example, a reference to *fly* in English language text might be disambiguated by references in nearby terms to airplanes and part of speech information indicating that the term *fly* is a verb, suggesting the meaning of "movement through the air," while occurrences of *fly* in the context of flying insects and bugs suggest another meaning. Natural language translation may also involve more difficult tasks, such as translating metaphors, sayings, or other more complex ideas from one culture to another. Sometimes the same effect is desired in both languages, so an emphatic statement or obscenity in a source language needs to be translated to cause the same reaction in the destination language.

While natural language translation is most often performed directly from one source language to the destination language, a more theoretically appeal-

ing method is to translate from the source language to a neutral, interlingual language and then from the neutral language to the destination language. In practice, there may be common languages or systems that serve as the *de facto* common language, whether it is a technical form of expression, such as algebraic notation, or a natural language that is commonly understood by source and destination communities of interest at a certain period, such as Latin, French, or English have been at periods in time.

The conversion of vocabulary terms, index terms, tags, or subject headings in information systems is a simpler process than translating full text from one natural language to another. There is not the recursive problem of working with larger and larger structures to ensuring that all are translated correctly. While the term *fly*, as described above, could have different meanings, the context can be viewed from a *set of terms* model rather than from the hierarchical model inherent in translating prose.

Converting all occurrences of one term to an occurrence of another term is far too simple a method for effective vocabulary conversion. More sophisticated methods addressing synonyms and homonyms are desirable if one is to capture enough of the differences in meanings in terminology.

The individual terms used by an individual can be shown to effect how they process information. In one study, native speakers of Russian, which has two words for blue, *goluboy* for lighter blues and *siniy* for darker blues, have been shown to differentiate between two blues faster when they belong to the two different categories than when both displayed blue figures were both *goluboy* or both *siniy* (Winawer, Witthoft, Frank, Wu, Wade, & Boroditsky, 2007). Native speakers of English discriminated between all the blue figures at the same rate, whether they were light blue or dark blue. This suggests that the terminology that people use does affect non-linguistic cognitive functions. Translation of the English language expression *the blue jacket* into Russian can be seen to present a problem; which blue was intended by the original author the the English language expression? The boundaries that exist in the cognitive structures of one individual may not be present in the native speakers of another language that has different boundaries. Translating both *goluboy* and *siniy* as *blue* when moving from Russian to English may be technically correct as a terminological translation, but the speakers of the two languages will clearly act differently. This reflects a basic limitation of natural language translation and simple vocabulary conversions.

A variety of techniques have been proposed for mapping terminology from one language system to another. This knowledge can then be represented using existing standards systems such as *Mapping SKOS* (Simple Knowledge Organization System). Sometimes equivalent vocabularies are sought, while in others, only a level of compatibility is desired (Vizine-Goetz, Hickey, Houghton,

& Thompson, 2004). Some work has focused on describing how vocabulary translations are made in existing systems. For example, Whitehead (1990) provides a largely descriptive work examining how the *Library of Congress Subject Headings* are mapped into the vocabulary used in the *Art and Architecture Thesaurus*. Doerr (2001) examines a wide range of techniques used in thesaurus mapping from an international perspective.

Some scholars have emphasized ways that systems can be improved (Olson & Strawn, 1997), some emphasizing how new distributed systems can integrate different vocabularies to provide an improved level of performance for searchers of linked systems (Nicholson, Dawson, & Shiri, 2000; Vizine-Goetz et al., 2004; Zeng & Chan, 2004), enriching services by including material derived from associations between features in multiple vocabularies.

The model proposed in the work below seeks to provide something like a mapping, but we explicitly focus on user-centered methods to provide representations that produce performance the same as or superior to the original representations.

4 Converting Uncontrolled to Controlled Vocabularies

Whether an information system should use a controlled vocabulary or whether it should allow the user to choose their own terminology is a decision that often faces system developers, purchasers, and users. Moreover, one is often faced with deciding, which vocabulary to use within a system. These decisions may be considered, moving from very simple situations to more complex situations.

When a system using the controlled vocabulary term x_c performs better than a system using the uncontrolled form of x_c, denoted as vocabulary term x_u, and where both have the same or similar meanings, we may compare the relative level of performance, and decide to use vocabulary term x_c instead of x_u, when

$$\mathcal{P}(x_c) > \mathcal{P}(x_u), \tag{1}$$

where $\mathcal{P}(x)$ represents the performance using term x.

As an example of this, assume that the performance associated with using the controlled vocabulary term *Subject headings* is worse than the performance associated with using the uncontrolled phrase *Controlled vocabularies* for the users of an information system. When Equation 1 does not hold, then one should not substitute the controlled vocabulary term for the uncontrolled term; the uncontrolled vocabulary term may be said to be superior or performatively equal in this context to the controlled vocabulary term.

A whole vocabulary system may be studied by examining the expected performance values of all the vocabulary in one system with the vocabulary in

another system. While this can be performed for any vocabulary system, we focus below on specific types of individual vocabulary conversions.

4.1 Converting Multiple Uncontrolled Terms to One Controlled Term

When should both of two uncontrolled terms be translated into a single corresponding controlled term? For example, when might two different informal and uncontrolled terms correspond to a single controlled term, such as when uncontrolled terms such as as *auto* or *car* might be changed to the controlled term *automotive*.

There will often be several uncontrolled terms corresponding to a single controlled term. The use of n_u uncontrolled terms $x_{u,1}, \ldots, x_{u,n_u-1}, x_{u,n_u}$ should be changed to the assignment of the controlled term x_c if and only if

$$\mathcal{P}(x_c) > \sum_{j=1}^{n_u} \Pr(x_{u,j})\mathcal{P}(x_{u,j}), \tag{2}$$

where $\Pr(x)$ is the probability of controlled term x. One should choose to use a controlled term x_c when the performance associated with its use is greater than the expected performance associated with the set of uncontrolled terms. The best single controlled term would be that controlled term whose use maximizes performance when all occurrences of the associated uncontrolled terms are replaced with the controlled term.

4.2 Ambiguous Uncontrolled Terms

Many uncontrolled terms are ambiguous, carrying more than a single word sense, and correspond to more than a single controlled term. Assume that there exists a single list or thesaurus of controlled vocabulary terms (Losee, 2007a). In the thesaurus, for example, the uncontrolled term *bank* may correspond to several controlled terms, such as terms associated with different meanings, including (1) *financial institutions*, (2) *the sloping edge of a river*, or (3) *to tilt an airplane during a turn*. Because there may be several possible controlled terms corresponding to the single uncontrolled term, a substitution decision rule for choosing possible vocabulary terms may take several forms.

One rule might be that one set of controlled terms is the set of controlled terms that maximizes the expected performance when substituted for the single uncontrolled term. A second rule would be to use a set of controlled terms that maximizes the expected performance when substituted for the single uncontrolled term and exceeds the uncontrolled term by performance amount z. This modifies a simple substitution rule to select the largest set of controlled

terms that, when substituted for the uncontrolled term, exceeds the uncontrolled term's performance by some fixed amount z. When $z = 0$, the largest set of controlled terms is obtained, while when z is some value greater than 0, one may obtain a more realistic situation where we want several controlled terms instead of the largest set of controlled terms available.

These rules may be formalized as suggesting that one should use the set of n_c controlled vocabulary terms $x_{c,1}, \ldots, x_{c,n_c-1}, x_{c,n_c}$ instead of uncontrolled term x_u if and only if

$$\left(\sum_{j=1}^{n_c} \Pr(x_{c,j}) \mathcal{P}(x_{c,j}) \right) + z > \mathcal{P}(x_u) \tag{3}$$

for a chosen value of z. One should choose a set of controlled terms that can cover all the needed variations. Selecting specific terms is discussed in the next section.

When developing a set of one or more controlled terms from a set of one or more uncontrolled terms, one may use a procedure that generates one or more controlled terms from each member of the set of uncontrolled terms, with the result being the union of the controlled terms being taken over all the sets of controlled terms that are produced, removing duplicate controlled terms. By comparing the expected performance of both source and destination sets, rational decisions may be made.

When the decision to convert one or more uncontrolled terms to one or more controlled terms is made, it can be based upon the expected performance values of both the sending and the receiving sets of terms. These may be a subset of a given vocabulary or may be the entire vocabulary. Using this allows us to both determine whether one should convert from one vocabulary to another vocabulary and whether, for a given situation, one vocabulary is better than another. Using predictive methods for computing the performance method $\mathcal{P}(x)$ (Losee, 2000), one can determine circumstances when one method or system would be better than another.

The application of these techniques, providing numeric examples and discussing their computation, is given in a later section.

5 Converting Controlled to Uncontrolled Vocabulary

Deciding to convert from controlled terminology to uncontrolled terminology is similar to deciding to move from uncontrolled terms to controlled terms. The primary difference is that there may be multiple uncontrolled destination terms associated with each single controlled source term, with each of these

uncontrolled terms representing the concept in question. In practice, individuals often move effortlessly from one uncontrolled vocabulary or sublanguage to another, such as when one uses terminology unfamiliar to the listener or reader, producing a term or statement that might be appropriate in one context but is unintelligible in another. Converting to an uncontrolled vocabulary from a single controlled term requires that the specific set of possible uncontrolled terms be specified before the conversion begins.

When converting from an uncontrolled term to a controlled term, the controlled term is often not a synonym of the uncontrolled term. The controlled term captures the context, including those uncontrolled terms that tend to co-occur. It may be the case that converting from a controlled term to an uncontrolled term is more likely to choose a synonym in the uncontrolled vocabulary than a related, co-occurring term in the uncontrolled vocabulary.

Substituting one or more uncontrolled terms in lieu of a controlled term x_c should occur only when

$$\mathcal{P}(x_c) \leq \sum_{i \ s.t. \ x_{u,i} \in s} \Pr(x_{u,i}) \mathcal{P}(x_{u,i}), \tag{4}$$

where s represents the set of terms co-occurring with the starter term $x_{u,k}$, the term in the destination uncontrolled vocabulary that has the highest performance $\mathcal{P}(x_{u,k})$ when substituted for the term x_c. There may be several different lists of co-occurring terms containing the starter term; s should be selected in this case so that the performance with list s substituting for the controlled term is superior to performance with any other list of co-occurring terms containing the seed term.

6 Conversion of Uncontrolled Terms to Other Uncontrolled Terms

Conversion from uncontrolled language or context U_1 to uncontrolled language or context U_2 can take place directly, or an intermediate controlled language C may be used so that the conversion goes $U_1 \rightarrow C \rightarrow U_2$. Converting directly from U_1 to U_2 has the advantage that no noise is added by mistakes made producing the intermediate version of the intended meaning or context, C.

On the other hand, producing an intermediate representation, C, has the advantage that error correction can occur as one moves from an imprecise term or slightly ambiguous term to a single controlled term. By transitioning through the controlled vocabulary, we can improve the accuracy of the representation.

For example, consider two controlled representations, one for A of binary feature frequencies 0000000 and one for B of 1111111. If the uncontrolled term with representation 0000001 occurs, we can guess that unless a very large error occurred, a small error of representing 0000000 as 0000001 occurred. This small error may be corrected, thus improving the process by providing a more accurate source from which the final uncontrolled term is derived. By using an intermediate term supporting error correction, the overall quality of the vocabulary conversion may be improved. Note that having poorly developed intermediate terminology that does not support the above form of error correction can result in an overall decrease in performance.

It is assumed here that using a controlled term allows for a perfect and unambiguous labeling of the concept; by converting to a perfect representation from an initially flawed uncontrolled term, and then producing another uncontrolled term from this second, perfect representation, the system avoids compounding errors. Consider an initial uncontrolled term that is 70% accurate, that is, 70% of the people using the term find the concept that they are looking for, and 30% are looking for another meaning or concept associated with the term. If this term were directly connected to one or more uncontrolled terms, errors in the processing may result in further degradation of the quality of the representation to below 70% accuracy. If, on the other hand, the initial term were "corrected" in its conversion to an appropriate unambiguous controlled term, the percent of errors in uncontrolled terms produced from this correct controlled term will be much better than from the uncontrolled terms produced directly from other uncontrolled terms. Using an intermediary controlled term improves performance, for example, when the sense of a term is determined by its context, and those terms with one sense have an appropriate controlled term assigned, and those with a second sense are assigned a different but appropriate controlled term, based on the different context in which it occurs. The controlled term may then be used in the production of an uncontrolled term from the second vocabulary system.

Also, as was mentioned above, the use of a single intermediate language results in far fewer language-to-language term maps than are necessary when direct translation from each possible source vocabulary to each possible destination vocabulary is used.

7 How to Compute Performance: An Example

Several different algorithms have been used to measure retrieval performance (Demartini & Mizzaro, 2006; Salton & McGill, 1983; Van Rijsbergen, 1979). Many were originally proposed for retrospective studies of retrieval experi-

ments, but some have also been developed for predictive applications as well as retrospective studies and some have emphasized the problems faced primarily by web search engines. The most common retrieval measures remain precision and recall, the percent of documents retrieved that are relevant and the percent of relevant documents that are retrieved, respectively. Single number measures of retrospective and predictive performance have been used in various studies (Borlund & Ingwersen, 1998; Jarvelin & Kekalainen, 2002; Losee, 2000; Voorhees, 2001). A measure is suggested here (Losee, 2007b) that is interpreted as the Percent of Perfect Performance (PPP), i.e., the degree to which the system has moved from random performance to the upper-bounds level of performance. The upper-bounds performance is the best achievable performance given the constraints of the model and the data under consideration, and methods for computing it are discussed elsewhere (Losee, 2007b). One begins by computing or estimating the Normalized Average Search Length (NASL), the average position of a relevant document in the ordered set of documents, scaled to be in the range of 0 to 1, so that 0 would be the best performance and 1 the worst. One may compute the PPP percent of perfect performance of system x as $\mathcal{P}(x) = \log(2 \cdot NASL_x)/\log(2 \cdot NASL_{Upperbounds_x})$.

The $NASL$ may be estimated or computed based on probabilistic considerations in some indexing situations. Given a single term being studied and assuming an optimal ranking algorithm, one may compute $NASL$ as $NASL = (1 - p + t)/2$, where p represents the probability that a relevant document has the feature in question, and t the unconditional probability that a document has the feature in question. When p increases and t decreases, we find that $NASL$ approaches 0, the best possible value, while when p decreases toward 0 and t approaches 1, then $NASL$ approaches 1, the worst possible value. Random ordering performance is indicated by $NASL = 0.5$.

Assume that we have a single term query and three documents, ordered so that the two documents with the single feature in question are first, with one of the two being relevant. The last of the three documents does not have the feature and is not relevant. These documents are ordered as in any realistic system presented with a term that is a good search term so that the documents with the term are ranked ahead of the documents without the term. One may assume that the query term occurs in relevant documents at a greater rate than in non-relevant documents; if this is not true, then the feature is recoded to its inverse, so, for example, *cats* would become *not cats*. Thus, we have a relevant document with the feature, a non-relevant document with the feature, and a non-relevant document without the feature, and documents occurring in that order. As a preliminary guess, we would expect that the first two documents both have the same feature, and thus the tied weight for both of these would place both conceptually at the center of the position of this set of two

documents, thus at $1/3$ of the way through all the documents. More formally, the $NASL$ is the expected position of the relevant document(s), which is at $1/3$ for this dataset. One may compute $NASL$ using the earlier formula by noting that p, the percent of relevant documents that have the feature, is 1 in this case, and t, the percent of all documents that have the feature, is $2/3$. Thus, $NASL = (1 - 1 + 2/3)/2 = 1/3$, as we intuitively argued.

Consider the following example of computing $NASL$ and PPP beginning with given parameters for two arbitrary systems, numbered 1 and 2. Assume the uncontrolled terms u_1 and u_2 with probabilities $p_1 = .9$, $p_2 = .8$, and for t values, $t_1 = .5$ and $t_2 = .6$. Thus $NASL_1 = (1 - .9 + .5)/2 = .3$ and $NASL_2 = (1 - .8 + .6)/2 = .4$. Assume that upper-bounds have an arbitrarily chosen $NASL = .05$. One may then compute $\mathcal{P}(1) = \log(2 \times .3)/\log(2 \times .05) = .2218 = 22.18\%$ and $\mathcal{P}(2) = \log(2 \times .4)/\log(2 \times .05) = 0.0969 = 9.69\%$. The PPP value provides the degree to which the method being examined has moved from random performance to the best possible performance, so our value of 22.18% indicates that the performance is about 22% of the way from random to upper-bounds performance.

The PPP for a single term may be computed as above, but what happens when two independent terms in a query are used together, each with possibly different PPP values? After converting each PPP value to a p value scaled from 0 to 1 by dividing by 100, instead of using 0% to 100%, the PPP value for two terms with p values denoted as p_1 and p_2, combined, will produce the PPP value $100(1 - (1 - p_1)(1 - p_2))$. Two PPP values, each of 50%, would give us $1-(1-.5)(1-.5)$ that results in $1-.25$ or 0.75, which, when multiplied by 100, produces 75%, indicating that two independent terms, each of which produce a PPP of 50% produce an improved PPP of 75% when used together.

Two extreme cases may also be useful to examine here. Two independent PPP values of 0% result in a combined or joint PPP of $100(1 - (1 - 0)(1 - 0)) = 0$ or 0%. This may be interpreted as consistent with two independent index terms, each of which performs at only a random level of performance (and contributes nothing toward the retrieval process) will, when combined, produce combined performance that is also random. Note that *when using traditional performance measures* such as precision, this methodology *will not work correctly* because the random level of precision in almost all systems is higher than 0. The case where two independent features are perfect, that is, they both have a PPP of 100% and are analyzed using the above formula, $100(1-(1-1)(1-1)) = 100(1 - 0) = 100$ resulting in a PPP of 100% as one would expect.

Table 1: Sample performance and probability figures for a source uncontrolled language and a destination uncontrolled language through an intermediate controlled language.

Uncontrolled Language 1	*Controlled Language*	*Uncontrolled Language 2*
$\mathcal{P}(u_1) = 22.18\%$,	$\mathcal{P}(c_1) = 20.00\%$	$\mathcal{P}(c_7) = 30.00\%$
$\Pr(u_1) = 0.5$,		$\Pr(c_7) = .90$
$\mathcal{P}(u_2) = 9.69\%$,		$\mathcal{P}(c_8) = 10.00\%$
$\Pr(u_2) = 0.5$,		$\Pr(c_8) = .09$
		$\mathcal{P}(c_9) = 1.00\%$
		$\Pr(c_9) = .01$

8 An Example of Vocabulary Conversion with Artificial Data

Hypothetical sample data is provided in Table 1. If the two features are equally probable, the average performance associated with using both features u_1 and u_2 is .1594, or 15.9% of the linear distance from random performance to the upper-bounds performance.

Assume that the performance associated with using controlled vocabulary term c_1 that is obtained when substituting it for terms u_1 and u_2 has $\mathcal{P}(c_1) =$ .2000. This data is captured using the same queries and documents as was used in determining the data for Uncontrolled Language 1 in the paragraph above. Using Equation 2 we can see that using the controlled term c_1 is superior to using the two uncontrolled terms with the expected performance computed above as .1594.

Given the expected values for three terms in Uncontrolled Language 2 (u_7, u_8, and u_9) as shown in Table 1, one can be justified using Equation 4 to substitute the set of terms u_7, u_8, and u_9 with the expected performance of .2791 in Uncontrolled Language 2 in place of controlled term c_1 with performance .2000.

One can convert from Uncontrolled Language 1 to Uncontrolled Language 2, with the controlled language as an intermediary. One could also directly convert from one uncontrolled language to another uncontrolled language, from Uncontrolled Language 1 to Uncontrolled Language 2 by comparing the expected performance for each of the uncontrolled languages.

9 Estimating Parameters

When human vocabulary developers propose a vocabulary, such as when scholars compile a thesaurus or an ontology, they are in effect conducting a Gedanken experiment in which they are imagining how different terms effect performance when used by a searcher (Cooper, 1978). Data may be gathered empirically too. Groups of searchers may be split so that half search a given database using one vocabulary system and half use another vocabulary system. Individuals may apply real or artificial queries to databases searchable using either one or the other of two vocabularies, allowing statistics for individual preferences for both vocabulary systems to be developed. For example, data gathered from users of an uncontrolled vocabulary and a given collection could be associated with the same queries being applied to a similar system using the controlled vocabulary. By collecting pairs of data, each member of the pair representing the use of a specific vocabulary, with the pair using the same query and the same collection, inequalities developed above that compare the performance of the two vocabularies may be used in decision making.

Where do the possible term pairs, such as in Table 1, come from when they are to be tested? Finding related terms has been examined in the information science literature for a variety of problems (Efthimiadis, 1996; Greenberg, 2001; Li & Yang, 2005; Lu & Keefer, 1995). One way to find related terms is to use dictionary entries, with the term being examined often being related to the other terms in the dictionary definition for the term. While some of the terms in a definition will have little relationship, for example, *the* or *an*, many terms will have a strong relationship and can serve as the basis for further analyses. Related terms can also be extracted from the sentences in which the seed term occurs, selecting those terms that co-occur with the seed term frequently and are substantial meaning-bearing terms beyond simple grammatical terms, e.g., *a, and, the*. Related terms in different languages may be extracted through the study of parallel documents in the two different languages, with a specific term in one language often occurring in text that matches with text in the other language containing matching terms (Li & Yang, 2005). Related terms may be filtered so that only those terms with matching parts-of-speech are selected, using automatic part-of-speech tagging programs, to increase the quality of the related terms being considered (Brill, 1994; Losee, 2001a; Murata, Ma, & Isahara, 2002). Term pairs may also be extracted from text by computing those terms that provide the highest information about the term in question. By using the expected mutual information measure, the information two terms provide about each other can be simply computed and used to measure the strength of association between terms (Church & Hanks, 1990; Li, 1989; Losee, 2001b).

One can learn which terms perform best through trial and error on various

randomly selected terms, or one can use related terms. For example, one could study what happens when the controlled term *cat* occurs in a document compared to performance when the document has related terms like *cats* or *feline*. One could also try studying the relationship between the controlled term *cat* and presumably unrelated uncontrolled terms such as *giraffe, oven*, or *influenza*.

The data about both vocabularies may be collected at almost the same time, or data may be collected in a much looser and asynchronous fashion. Information needs may be understood as lasting over longer chronological periods and the accesses made over time to identical or highly similar documents may be recorded.

In many circumstances, the data needed to make the above calculations is not readily available. Instead, the data, most often relevance-related data, must be estimated from user's histories, the history of others, or general users' behaviors. Supervised learning can be used in decision making, with all the relevance data provided being used in estimating parameters (Knill & Young, 1997; Duda, Hart, & Stork, 2001). More recently, people have argued for the use of semi-supervised learning, in which some supervised learning is available, but learning takes place even when not all data is labeled as to relevance (Chapelle, Scholkopf, & Zien, 2006). For example, given a few of the best relevant and non-relevant documents, one can infer the underlying parameters of both classes of documents (Losee, 1987). Relevance judgments may be inferred from various user behaviors when relevance judgments are not explicitly provided (Kelly, 2005).

Parameters may also be computed using widely accepted relationships involving term rank, probability, discrimination, and performance. A term's rank (e.g. most popular, second most popular, third most popular, etc.) may be computed from the dataset being used. Given the term rank, and thus the probability from Zipf's Law, and Luhn's suggested distribution describing the discrimination power of terms (Egghe, 1999; Losee, 2001b; Luhn, 1958; Pao, 1978), as well as formula describing discrimination (Duda et al., 2001; Losee, 1988), one may algebraically solve for the parameters of the relevant documents, given that the parameters of non-relevant documents may be estimated from the parameters of all documents (Croft & Harper, 1979). The discrimination power may be inferred to be constant for all terms, or to increase for mid-frequency terms, as suggested by Luhn (1958) and Pao (1978), or the shape of the discrimination vs. rank curve may be inferred from those available terms where some supervised learning or explicit or implicit feedback is available. The parameters may be estimated by a user or subject specialist (Cooper, 1978) and then used in decision making. One may conversely estimate ranges of parameter values from the historical decisions that were made by specialists.

10 Example: Performance with User-supplied Tags from Flickr.com

What data do we find in existing systems when people provide their own uncontrolled tags to objects? Using a variety of tags to represent a category, or portions of a category, can result in similar items being spread over a range of different terms that might be the terms used in searching. Searching using one search term would retrieve objects with that term, but similar objects that were tagged with a different, but synonymous search term, may remain undiscovered.

Below we empirically analyze some data from the collection of images at the website *http://flickr.com* as it existed in June 2007. The number of images is approximately 100 million. Containing millions of images submitted by users for viewing by themselves and others, some images have free text descriptions; they can also be assigned tags generated by the image submitter. For many of the analyses, we focus on retrieving images based only on the assigned tags. These terms and the associated statistics represent the type of terms used by people who are not information professionals.

The probabilities used in the formulas above may be computed from the number of images having the appropriate tags. The t for a specific tag is computed based on the percent of all images that have the term in question. The p for a specific tag is computed as the percent of the set of all relevant images that have the tag in question. The number of (relevant) images with tags for *cat* and *cats* is approximated by the size of the returned set from the query `cat OR cats OR kitten OR kittens OR feline OR felines`. The number of relevant images may be estimated with the feature as the number retrieved with the query, e.g., `cat`. There will be non-relevant images in both groups of images, but we treat both groups as having non-relevant images at the same rate, and thus estimate p by dividing the number of images with `cat` divided by number in the larger set of relevant images. For terms related to younger cats, we treat the relevant images as those having the terms *kitten* or *kittens* in the images' tags. Note that this example simplifies the problem for illustrative purposes, with a few remaining pictures about cats but using other terminology not being included here, and some other pictures, e.g., about caterpillars, being included because of their use of the term *cat*.

Most of the images retrieved using these terms are of the appropriate category of cats. Some terms that retrieve many other images are omitted. Terms from other languages (e.g. the French term *chat*) are omitted, limiting this study and the probabilities to English language terms only.

Table 2 provides term frequencies, the associated probabilities and performance figures. Using Equation 2, we can see that using the controlled vocabu-

Table 2: Performance for variants of the term *cat*. The last line contains possible performance figure for a controlled term denoted as *CAT* that could be used whenever uncontrolled terms *cat, cats, feline*, or *felines* are used, with appropriate term frequencies. Upper bounds performance is treated arbitrarily as 0.005. All whole numbers represent thousands of flickr.com images. Note that the number of significant digits shown is large due to the wide range of probabilities presented.

	Relevant Images			*All Images*			
Tag	*Tag Present*	*All*	p	*Tag Present*	*All*	t	$\mathcal{P}$
cat	1065	1454	0.732462	1065	100000	0.01065	27.7831%
cats	503	1454	0.345942	503	100000	0.00503	9.05284%
feline	54	1454	0.0371389	54	100000	0.00054	0.809644%
felines	8	1454	0.00550206	8	100000	0.00008	0.118059%
kitten	174	412	0.42233	174	100000	0.00174	11.8507%
kittens	82	412	0.199029	82	100000	0.00082	4.79694%
CAT	1454	1454	1.	1454	100000	0.01454	91.8718%

lary term *CAT* (assigned to an image when *cat*, *cats*, *feline* or *felines* is present) produces superior performance to that obtained with the uncontrolled vocabulary, regardless of the probabilities associated with the uncontrolled terms. If one were to assume that each specific type of cat were labeled by the appropriate tag type if and only if the image would be of interest to those searching for images of this type, then one could perform an analysis as above with the various tags.

Converting terms when some are very specific and some are more general is problematic. One can easily model the performance with narrower or broader terms in a hierarchical relationship (Doerr, 2001; Losee, 2007a). Determining which specific narrower term should be used may be virtually impossible without additional contextual knowledge. For example, an image tagged with the term *cat* may actually be a picture of a female or a male cat, or it might be one of a type of cat, such as a *Maine Coon* cat, a *Norwegian Forest* cat, or an *Abyssinian* cat. All of these are tags used in flickr.com, but often these types of cats are labeled by submitters as just *cat*. For example, the author does not ever recollect describing one of his cats to coworkers or students as anything but a *cat*, although the cat once was described as a fine example of a *tortie* cat by a veterinarian.

In many instances, a term may have multiple senses (a homonym) and be considered ambiguous in many situations. This differs from what occurs with the concept cats, in general, and specific types of cats, where terms may

be better described as *general* terms rather than *ambiguous* terms. In the example given earlier, we considered the case where the term *bank* can have several different senses. One may locate many images associated with the different meanings of the term *bank* in flickr.com by using searches such as `bank AND (airplane OR plane OR jet)` or `bank AND (river OR stream)` or `bank AND (financial OR money).` Clearly, the rules developed early in this article may be applied to compute performance parameters, given empirical data such as that derived from flickr.com.

11 Rules

From this type of empirical data, where we have several terms or tags that are essentially synonyms, we can test rules. One may also make general analytic rules about when to use controlled or uncontrolled terms. One simple rule about when to use controlled terms instead of uncontrolled terms is as follows:

> **Controlled Vocabulary Superiority in Synonyms.** *Given a controlled term and a synonymous set of uncontrolled terms, and each of the synonymous uncontrolled terms only has the meaning associated with the controlled term, then using the controlled vocabulary term will always perform better than or equal to the expected performance obtained with using the individual uncontrolled terms.*

While the documents with the uncontrolled term are retrieved first, other relevant documents (without the term) will be spread thorough the remaining documents in the database. As one can see in Table 2, all the individual uncontrolled terms that mean the same thing as CAT (*cat, cats, feline,* and *felines*) have a joint frequency that is the same as the total frequency for CAT, producing a p value of 1 for CAT, and CAT has a performance figure for the controlled term that exceeds the performance for each of the individual uncontrolled performance values. Equation 2 holds for all possible performance values for the uncontrolled terms, regardless of their relative probabilities. A simple calculation of this equation was tried for a range of values and the Equation held in all cases, as one might expect. The use of the PPP measure assumes retrieval of all the relevant documents, and is informally a high recall measure.

In other instances, a single term may have multiple meanings. *Homonyms* such as these can clearly benefit from having labels that better capture their meaning in an unambiguous manner. A rule addressing the application of controlled vocabulary to such terms is:

> **Controlled Vocabulary Superiority to Uncontrolled Homonyms.** *When a single uncontrolled term has multiple senses, and each sense is*

only represented by that particular uncontrolled term, then assigning an appropriate controlled term to terms of each term sense (based on the context) will provide performance greater than or equal to that obtained with the uncontrolled term.

It is difficult to compare the controlled and uncontrolled vocabulary here because the number of images or documents drops as one moves to the smaller numbers for each of the controlled vocabulary terms from the single uncontrolled vocabulary term. If there were 3 different meanings for an uncontrolled term, evenly distributed, and there were 99 documents or images, assigning a meaning-specific controlled term to each document would result in 33 documents for each of the 3 controlled terms. When the number being retrieved is constant, the performance measure cannot be used directly to compare the use of controlled and uncontrolled vocabulary when comparing unequal numbers of images or documents. If we assume that all the 33 documents with a given controlled term are relevant to its corresponding concept, then the average position of a relevant document would be at $33/2$, while when working with the 99 documents with the uncontrolled term, with all documents appearing equal in ranking due to all having the same terms, then the relevant documents would be spread evenly throughout this set and the average position for a relevant document would be at $99/2$, a much longer search than would be the case with the average relevant document position at $33/2$ that occurs when using the controlled vocabulary term. This can be generalized to any number of relevant and non-relevant documents.

This type of argument provides definitive statements about information systems performance values that can only be approximated by qualitative or experimental work (Losee, 1998). While formal proofs of rules such as these can be developed, these rules can also be studied experimentally, as one often sees in the indexing literature (Cleveland & Cleveland, 2001; Cleverdon, 1967; Foskett, 1996; Salton & Lesk, 1968).

12 Discussion and Conclusion

Designers of information systems often choose for the systems to represent items using either a controlled vocabulary system or uncontrolled vocabularies. When using metadata to represent the characteristics of a system, a particular system must be chosen and a set of values selected to represent items in the collection. Moving from one representation system to another is reasonable if and only if overall system performance does not degrade when the change is made, costs being held constant for all systems. Above we developed procedures that allow one to decide when these vocabulary conversions

are appropriate and when they are ineffective or harmful.

The above models provide some advantages that seem uncommon in studying language relationships. For example,

- this model explicitly provides an equal or superior vocabulary, and, more generally, provides a measure of the quality of a specific term translation;
- the PPP (Percent Perfect Performance) measure may be predicted analytically, enabling one to show performance of using terms as the term parameters vary;
- the PPP measure is explicitly based on individuals relevance judgments, and thus on subjective user values and preferences and is truly user-centered;
- the PPP measure may vary over time as a the values or preferences of an individual, group, or organization changes over time, allowing this model to be used to produce a vocabulary that reflects the past, present, or anticipated user needs of any individual, group, or organization;
- PPP measures may be combined for use with multiple terms; and
- the PPP measure may be computed in distributed environments (Losee & Church, 2004) and may be used for distributed vocabulary systems.

Conversions may occur directly from one vocabulary to another or may move through a third, intermediate vocabulary. Using an intermediate language allows for the possibility of error correction to occur as terms that are "near" a term in the intermediate language can be moved to this term, standardizing the concepts and ideas used during the conversion process. The number of sets of parameters and conversion rules is smaller when using an intermediate language than the number of sets needed when the possibility of converting directly from any source language to any other destination language. The nature of this intermediate language and its possible degree of perfection and cultural and domain dependence are certainly areas where further study is needed.

The methods for converting above may be used to convert individual sets of terms, or to convert an entire vocabulary. The decision to convert one or two terms to another small set of terms is relatively easy. Converting an entire vocabulary is more difficult. Clearly, if one designs a new vocabulary, there may be flaws in it, and applying the term by term decisions may suggest that most, but not all, of the terms in the older language should be converted into the planned terms in the new language. Instead, one may compute or predict the expected performance of the entire vocabulary, as suggested at the end of

Section 3, and decide to use the vocabulary with the highest expected performance.

An open question is whether terms that produce the best retrieval for a person, group, or organization also serve as the best features for ordering documents for browsing. Arranging either structured or unstructured data or a mixture of both for browsing can be done optimally based upon features used (Losee, 2006a). Because the methods for choosing optimal ordering are different for optimal ordering for browsing and optimal ordering for retrieval, the features that are best for end users may differ from one method to another, but the nature of this difference is not well understood.

References

Aronoff, M., & Fudeman, K. (2005). *What is Morphology?* Blackwell, Malden, MA.

Borlund, P., & Ingwersen, P. (1998). Measures of relative performance and ranked half-life: Performance indicators for interactive IR. In Croft, W. B., Moffat, A., Van Rijsbergen, C. J., Wilkinson, R., & Zobel, J. (Eds.), *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Melbourne, Australia*, pp. 324–331. ACM Press, New York.

Bowker, L., & Hawkins, S. (2006). Variation in the organization of medical terms. *Terminology*, *12*(1), 79–110.

Brill, E. (1994). Some advances in transformation-based part of speech tagging. In *Proceedings of the Twelfth National Conference on Artificial Intelligence (AAAI-94)*, pp. 722–727 Menlo Park, CA. AAAI Press.

Chapelle, O., Scholkopf, B., & Zien, A. (2006). *Semi-Supervised Learning*. MIT Press, Cambridge, Mass.

Chomsky, N. (1965). *Aspects of the Theory of Syntax*. MIT Press, Cambridge, Mass.

Church, K. W., Gale, W., Hanks, P., Hindle, D., & Moon, R. (1994). Lexical substitutability. In *Computational Approaches to the Lexicon*, pp. 153–177. Oxford U. Press, Oxford.

Church, K. W., & Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational Linguistics*, *16*(1), 22–29.

Cleveland, D. B., & Cleveland, A. D. (2001). *Introduction to Indexing and Abstracting* (Third edition edition). Libraries Unlimited, Englewood, Colo.

Cleverdon, C. (1967). The Cranfield tests on index language devices. *Aslib Proceedings*, *30*, 172–181.

Cooper, W. S. (1978). Indexing documents by Gedanken experimentation. *Journal of the American Society for Information Science*, *29*(3), 107–119.

Croft, W. B., & Harper, D. (1979). Using probabilistic models of document retrieval without relevance information. *Journal of Documentation*, *35*(4), 285–295.

Demartini, G., & Mizzaro, S. (2006). A classification of IR effectiveness metrics. In *Advances in Information Retrieval: Lecture Notes in Computer Science, 3936*, pp. 488–491. Springer-Verlag, Berlin.

Dodd, B., Campbell, R., & Worrall, L. (1996). *Evaluating Theories of Language: Evidence from Disordered Communication*. Singular Publishing Group, San Diego, CA.

Doerr, M. (2001). Semantic problems of thesaurus mapping. *Journal of Digital Information*, *1*(8). http://jodi.tamu.edu/Articles/v01/i08/Doerr/.

Duda, R. O., Hart, P. E., & Stork, D. G. (2001). *Pattern Classification* (2nd edition). Wiley, New York.

Efthimiadis, E. N. (1996). Query expansion. In *Annual Review of Information Science and Technology*, pp. 121–187. Information Today, Inc., Medford, NJ.

Egghe, L. (1999). On the law of Zipf-Mandelbrot for multi-word phrases. *Journal of the American Society for Information Science*, *50*(3), 233–241.

Foskett, A. C. (1996). *The Subject Approach to Information* (Fifth edition). Library Association Pub., London.

Furnas, G. (1985). Experience with an adaptive indexing scheme. In *Proceedings of the ACM CHI*, pp. 131–135. ACM.

Greenberg, J. (2001). Optimal query expansion (QE) processing methods with semantically encoded structured thesauri terminology. *Journal of the American Society for Information Science and Technology*, *52*(6), 487–498.

Grefenstette, G. (1992). Use of syntactic context to produce term association lists for text retrieval. In *Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Stockholm, Sweden*, pp. 89–97 New York. ACM Press.

Jarvelin, K., & Kekalainen, J. (2002). Cumulative gain-based evaluation of IR techniques. *ACM Transactions on Information Systems*, *20*(4), 422–446.

Kelly, D. (2005). Implicit feedback: Using behavior to infer relevance. In Spink, A., & Cole, C. (Eds.), *New Directions in Cognitive Information Retrieval*, pp. 169–186. Springer Publishing, Netherlands.

Knill, K., & Young, S. (1997). Hidden Markov models in speech and language processing. In Young, S., & Bloothooft, G. (Eds.), *Corpus-Based Methods in Language and Speech Processing*, pp. 27–68. Kluwer, Dordrecht.

Li, K. W., & Yang, C. C. (2005). Automatic crosslingual thesaurus generated from the Hong Kong SAR Police Department web corpus for crime analysis. *Journal of the American Society for Information Science and Technology*, *56*(3), 272–282.

Li, W. (1989). Mutual information functions of natural language texts. Tech. rep. 89–008, Santa Fe Institute, New Mexico.

Losee, R. M. (1987). The effect of database size on document retrieval: Random and best-first retrieval models. In *ACM Annual Conference on Research and Development in Information Retrieval*, pp. 164–169 New York. ACM Press.

Losee, R. M. (1988). Parameter estimation for probabilistic document retrieval models. *Journal of the American Society for Information Science*, *39*(1), 8–16.

Losee, R. M. (1998). *Text Retrieval and Filtering: Analytic Models of Performance*. Kluwer, Boston.

Losee, R. M. (2000). When information retrieval measures agree about the relative quality of document rankings. *Journal of the American Society for Information Science*, *51*(9), 834–840.

Losee, R. M. (2001a). Natural language processing in support of decision-making: Phrases and part-of-speech tagging. *Information Processing and Management*, *37*(6), 769–787.

Losee, R. M. (2001b). Term dependence: A basis for Luhn and Zipf models. *Journal of the American Society for Information Science*, *52*(12), 1019–1025.

Losee, R. M. (2006a). Browsing mixed structured and unstructured documents. *Information Processing and Management*, *42*(2), 440–452.

Losee, R. M. (2006b). Is 1 noun worth 2 adjectives? Measuring the relative feature utility. *Information Processing and Management*, *42*(5), 1248–1259.

Losee, R. M. (2007a). Decisions in thesaurus construction and use. *Information Processing and Management*, *43*(4), 958–968.

Losee, R. M. (2007b). Percent perfect performance (PPP). *Information Processing and Management*, *43*(4), 1020–1029.

Losee, R. M., & Church, L. (2004). Information retrieval with distributed databases: Analytic models of performance. *IEEE Transactions on Parallel and Distributed Systems*, *15*(1), 18–27.

Lu, X. A., & Keefer, R. B. (1995). Query expansion/reduction and its impact on retrieval effectiveness. In *Overview of the Third Text REtrieval Conference (TREC-3)*, pp. 231–239. National Institute of Standard and Technology, Computer Systems Laboratory, Gaithersburg, MD.

Luhn, H. (1958). The automatic creation of literature abstracts. *IBM Journal of Research and Development*, *2*(2), 159–165. The article is also included in *H. P. Luhn: Pioneer of Information Science, Selected Works*.

Maniez, J. (1997). Database merging and the compatibility of indexing languages. *Knowledge Organization*, *24*(4), 213–223.

Manning, C. D., & Schutze, H. (1999). *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, Mass.

Murata, M., Ma, Q., & Isahara, H. (2002). Comparison of three machine-learning methods for Thai part-of-speech tagging. *ACM Transactions on Asian Language Information Processing*, *1*(2), 145–158.

Newmeyer, F. J. (1986). *Linguistic Theory in America* (Second edition). Academic Press, New York.

Nicholson, D., Dawson, A., & Shiri, A. (2000). HILT: A pilot terminology mapping service with a DDC spine. *Cataloging & Classification Quarterly*, *42*(3/4), 187–200.

Olson, T., & Strawn, G. (1997). Mapping the LCSH and MeSH systems. *Information Technology and Libraries*, *16*(1), 5–19.

Pao, M. L. (1978). Automatic text analysis based on transition phenomena of word occurrences. *Journal of the American Society for Information Science*, *29*(3), 121–124.

Pearce, D. (2002). A comparative evaluation of collocation extraction techniques. In *Third International Conference on Language Resources and Evaluation*, Vol. 5, pp. 1530–1536 Las Palmas, Canary Islands, Spain.

Salton, G., & Lesk, M. (1968). Computer evaluation of indexing and text processing. *Journal of the ACM*, *15*(1), 8–36.

Salton, G., & McGill, M. (1983). *Introduction to Modern Information Retrieval*. McGraw-Hill, New York.

Van Rijsbergen, C. (1979). *Information Retrieval* (Second edition). Butterworths, London.

Vivaldi, J., & Rodriguez, H. (2007). Evaluation of terms and term extraction systems. *Terminology*, *13*(2), 225–248.

Vizine-Goetz, D., Hickey, C., Houghton, A., & Thompson, R. (2004). Vocabulary mapping for terminology services. *Journal of Digital Information*, *4*(4). http://jodi.tamu.edu/Articles/v04/i04/Vizine-Goetz/.

Voorhees, E. M. (2001). Evaluation by highly relevant documents. In Kraft, D. H., Croft, W. B., Harper, D. J., & Zobel, J. (Eds.), *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, New Orleans, Louisiana*, pp. 74–82 New York. ACM Press.

Weeds, J., & Weir, D. (2005). Co-occurrence retrieval: A flexible framework for lexical distributional similarity. *Computational Linguistics*, *31*(4), 439–475.

Whitehead, C. (1990). Mapping LCSH into thesauri: The AAT model. In *Beyond the Book: Extending MARC for Subject Acccess*, pp. 81–96. G. K. Hall and Co., Boston.

Winawer, J., Witthoft, N., Frank, M. C., Wu, L., Wade, A. R., & Boroditsky, L. (2007). Russian blues reveal effects of language on color discrimination. *Proceedings of the National Academy of Science, USA*, *104*(19), 7780–7785.

Zeng, M., & Chan, L. (2004). Trends and issues in establishing interoperability among knowledge organization systems. *Journal of the American Society for Information Science and Technology*, *55*(5), 377–395.